DOCUMENT RESUME

ED 183 130 IR 008 026

AUTHOR Thomas, David B.
TITLE Effects of Alternative Response Mode Requirements on the Acccuracy of Student Response in Computer-Based Testing. Report Series 1979.
INSTITUTION Iowa Univ., Iowa City. Computer Center.
PUB DATE 79
NOTE 8p.
AVAILABLE FROM CAI Laboratory, University of Iowa, Weeg Computing Center, Iowa City, IA 52242

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students: *Computer Assisted Instruction: *Educational Research: Reaction Time: *Response Mode: Response Style (Tests): Typewriting: Typists

ABSTRACT

This study was designed to determine the accuracy with which a student in a computer-based testing situation will be able to accurately communicate the selected answer to the computer. Such a test was administered to 34 students, with the answer to each item supplied. Examinees, who were identified as touch typists or non-typists, used either the numbers 1 through 5 or the letters A through E as possible answers. Test scores were significantly higher for examinees taking the test under the letter condition, and typists in this group completed the task more quickly: the non-typists worked faster under the number condition. An inspection of the errors revealed that adjacent and blank key errors occurred more frequently for the number condition and for nontypists. The limitations of this study which affect the generalizability of the findings are discussed, and it is recommended that the study be replicated with more realistic looking items and a more balanced sample.
(Author/JEG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183130

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EFFECTS OF ALTERNATIVE RESPONSE MODE REQUIREMENTS ON THE ACCURACY OF STUDENT RESPONSE IN COMPUTER-BASED TESTING

By
David B. Thomas
Computer Assisted Instruction Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
University of Iowa
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Weeg Computing Center
The University of Iowa
Report Series 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Bobby R. Brown
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR000026

EFFECTS OF ALTERNATIVE RESPONSE MODE REQUIREMENTS ON THE ACCURACY OF STUDENT RESPONSE IN COMPUTER - BASED TESTING

ABSTRACT

A computer-based test was administered to 34 students. The answer to each item was given along with the item. Examinees used either the numbers 1 through 5 or the letters A through E as possible answers. Each examinee was identified as a touch typist or a nontypist.

Total scores were significantly higher for examinees taking the test under the letter condition. Typing ability had no general affect. A significant interaction ($p < .10$) was found involving time to completion of the test. Typists under the letter condition completed the task more quickly. Nontypists were faster under the number condition.

An inspection of the errors revealed that adjacent and blank key errors occurred more frequently for the number condition and for nontypists.

INTRODUCTION

Computer-supported instruction takes many forms, from simple drills and tutorials through complex simulations to computer-managed instruction. One form of computer support of instruction is the printing and grading of tests administered off-line. A growing application is the adminstration of tests at computer terminals where immediate scoring and within-test adaptation of various forms may occur. Inasmuch as computer-based testing may soon be widespread, it is imperative that research on the nature of this form of testing be undertaken so that the advantages and disadvantages of this method of testing may be made known to users.

The study described in this report focuses on the nature of the interaction between an examinee and a computer terminal within the computer-based testing situation. The present study relates to the accuracy with which a student who knows (or has determined what s/he believes to be) the answer to an item will be able to accurately communicate that answer to the computer. Previous research suggests that this is not a straightforward matter. Accuracy seems to be influenced by the nature of the answer response medium.

Test scores are affected by the form of a machine-scorable answer sheet (Miller and Minor, 1963). The specific format differences tested by these authors were horizontally-and vertically-oriented response mark delimiters with response cues (i.e., 1, 2, 3, 4, 5) placed over, under, or alongside the dilimiters. The stimulus materials consisted of an item number with a corresponding "answer" numeral beside it. The examinee (fourth graders, eighth graders, and college students) was to mark the "answer" on the mark sense answer sheet. At each grade level, difference in answer forms were significant at the .01 level.

The Miller and Minor study described was performed using a separate answer sheet of varying format. Other researchers have employed both separate answer sheets (SAS) and answers recorded directly in the test booklet (AIB). elementary school pupils (grades 1-3) were studied by Cashen and Ramseyer (1969) who found performance levels differed less as grade level increased, but that only the third grade examinees showed no treatment difference. Significant accuracy differences between AIB and SAS scores were shown for grades 3, 4, and 6 in favor of the AIB condition by Muller, Calhoun, and Orling (1972). Significant variance differences (lower for AIB) were also shown by these authors.

These studies, conducted with elementary school children, suggest that separate criteria may be required when differently styled answer sheets are employed. The test items used were such that the answer was either given or obvious to most pupils. Thus the conclusion may be drawn that the performance differences may be attributed to short term memory, answer sheet marking errors, or similar factors other than knowledge of the correct answer.

Transferring an answer choice to a computer terminal presents yet another dimension to the response mode question. A computer terminal may be considered more demanding by some examinees (more "technical") or the array of keys may be considered difficult to use (answer codes are essentially randomly placed).

Adults were studied by Conrad and Hull (1968) in a task involving two alternative numeric keysets. One was a standard adding machine keyset with keys 1, 2, 3 on the bottom row; the other keyset was similar to a touch-tone telephone with 1, 2, 3 on the top row keys. Each participant (housewives, supposedly untrained in the task) entered eight-digit codes for 30 minutes on each of four days. No speed differences were detected between the two groups. For whole codes (all eight digits) and for individual digits, the telephone style keyset showed fewer errors and fewer uncorrected errors. These data show clearly that the keyboard format has an influence on the accuracy of response.

The studies referenced above suggest many factors other than knowledge which could influence performance on a computer-based test. Alternative keyboards, as suggested indirectly by Miller and Minor (1963) and directly by Conrad and Hull (1968) may decrease measured performance; the unfamiliar response input device itself may cause a problem as suggested by the results reported by Cashen and Ramseyer (1969) and Muller, Calhoun, and Orling (1972). The present study attempts to determine whether responses of A, B, C, D, or E, which are scattered on the typewriter-like keyboard, show lower levels of accuracy than responses employing numerals, which are located in a single line near the top of the keyboard.

MATERIALS

Experimental tests were prepared in two forms. The first form used as stimuli the letters A, B, C, D, and E. The second form used the numbers 1, 2, 3, 4, and 5. Manual and computer-based versions of the tests were employed. Both versions consisted of 125 "items" in which the examinee was presented with a character (letter A-B-C-D-E or number 1-2-3-4-5) and asked to enter that character as the "answer" on the response medium. The response medium for the manual version was a standard mark-sense answer sheet; for the computer-based version the response medium was a typewriter-like keyboard.

The manual test was presented via a booklet containing 11 items per page. Each examinee was instructed to locate the answer, fill in the appropriate block on the answer sheet, then to turn the test page for the next item. This procedure was repeated until all items had been responded to.

The computer-based version was presented by way of a Super Beehive CRT terminal connected to an HP 2000 computer system. Items were presented with one item visible at a time. The examinee was instructed to press the character on the keyboard corresponding to the "answer" displayed on the screen, then to depress the "RETURN" key to send the response to the computer. The next item was then displayed automatically. For the first 20 items only, corrective feedback was displayed whenever the examinee responded with a character other than a member of the expected response set. In that event, a new response was requested.

The manual and computer versions were identical in content and in item order. Each test contained an equal number of alternative stimuli, i.e., either 25 each A, B, C, D, E or 25 each 1, 2, 3, 4, 5. The stimuli were randomly ordered with the restriction that there be no adjacent item duplication (e.g., E-E or 2-2). The stimulus order was identical in computer and manual versions. The alternative test forms (letters-numbers) were designed to be comparable, e.g., the number form was directly derived from the letter form. Each "A" was converted to a "1", each "B" to a "2" and so on. This conversion was made for both the computer and manual versions of the instrument. Four test forms resulted, a manual and a computer version for both letter and number forms.

PROCEDURE

Subjects — Participants were randomly selected from an upper-level educational media class in the College of Education at the University of Iowa. Attendance at a "CAI orientation" session was mandatory for all students. The experimental materials were presented prior to the mandatory session. Approximately one-third (N = 45) were asked to participate in this study. No students refused to participate.

Method — Participants were randomly assigned to either a number (N) or a letter (L) condition. Sample sizes of 24 (N) and 21 (L) were obtained. The appropriate manual version of the test was then administered to each examinee. When the examinee completed the manual version, a proctor prepared the terminal for the computer-based version. Directions were presented on the CRT screen. Each participant was asked to enter the ID number provided, his/her gender, and whether s/he could touch type.

Item responses and the total time to complete the test were retained for later analysis. At the conclusion of the session, each participant was debriefed, then presented with the mandatory CAI orientation.

RESULTS

Forty-five participants completed the manual instrument. The response sets for three of these were eliminated from further analysis because of either random or incomplete responses. Responses on the computer instrument were considered to be usable for these participants. Eleven sets of computer responses (4 letter [L], 7 number [N]) were not used in the analyses. Ten of these sets were lost due to a computer failure; the remaining set of responses (N condition) was rejected because the low score (80) was judged invalid. This score was 27 points less than the next lowest score. Thus a total of 42 response sets (20-L, 22-N) were available for the manual test and 34 response sets (17-L, 17-N) were available for the computer based test.

Analysis of the manual test scores indicated a very high rate of accuracy for both the N and L versions. Means for both conditions exceeded 124 out of 125 items. Score variance was much higher for the letter condition. Table 1 indicates the means and standard deviations for the manual test scores.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR MANUAL TEST SCORES

Condition	N	Mean	S.D.
Numbers	22	124.55	1.06
Letters	20	124.80	.52

Score means and standard deviations for the computer-based test are reported in Table 2. The condition by typing ability design was analyzed by way of an unweighted means analysis of variance (Glass and Stanley, 1970). The results of these analyses are reported in Table 3. The analysis of total score indicated a difference ($p < .05$) between the scores obtained under the number and letter stimuli. The mean number condition score was 119.94 compared to a letter condition mean of 123.29, a difference of more than three points. A similar result was found when the first 25 items were analyzed separately. Participants tested under the letter condition obtained a mean score of 24.82 compared to 23.47 for those tested under the number condition, a difference of nearly one and one-half points. No significant difference was found beyond the first 25 items.

TABLE 2

MEAN AND STANDARD DEVIATIONS OF COMPUTER SCORES AND TIMES TO COMPLETION FOR ALL CELLS

		Letters			Numbers			Combined		
		Mean	SD	N	Mean	SD	N	Mean	SD	N
Total Computer Score	Type	123.20	2.04	10	120.75	2.77	4	122.50	2.53	14
	No Type	123.43	1.82	7	119.69	4.96	13	121.00	4.29	20
	Combined	123.29	1.99	17	119.94	4.29	17	121.62	3.74	34
1-25 Computer Score	Type	24.80	.40	10	23.50	2.60	4	24.43	1.55	14
	No Type	24.86	.35	7	23.46	2.00	13	23.95	1.75	20
	Combined	24.82	.38	17	23.47	2.15	17	24.15	1.68	34
26-125 Computer Score	Type	98.40	1.80	10	97.25	2.49	4	98.07	2.09	14
	No Type	98.57	1.76	7	96.23	4.24	13	97.05	3.72	20
	Combined	98.47	1.75	17	96.47	3.91	17	97.47	3.19	34
Total Computer Time	Type	193.90	18.77	10	222.00	53.18	4	201.93	34.94	14
	No Type	236.29	34.95	7	204.31	42.28	13	215.50	42.70	20
	Combined	211.35	33.85	17	208.47	45.71	17	209.91	40.25	34

TABLE 3

SIGNIFICANT F RATIOS

Variable	Factor	F	P
Total computer score	stimulus	5.354	<.05
1-25 computer score	stimulus	4.757	<.05
Total computer time	stimulus x typing skill	4.074	<.10

The time required to complete the test was also analyzed by the method of unweighted means. An interaction ($p < .10$) between test stimulus and typing skill was found. Typists took less time and had smaller variance under the letter condition, Nontypists took less time under the number condition. The obtained distribution of typists and nontypists within the N and L conditions was uneven. A chi-square computed for the obtained frequencies was significant at the ($p < .05$) level, indicating that the typists were not evenly distributed in the N and L conditions. The number condition consisted predominantly of nontypists; the letter condition was more or less equally composed of typists and nontypists.

Errors were classified into one of three categories: (1) an incorrect choice is made from among the alternatives (e.g., typing B, C, D, or E when A is correct); (2) an incorrect key is pressed (e.g.. typing S when A is correct), referred to as an "adjacency error" when the struck key is within one key in distance from the correct answer, or "target" key; and (3) a blank is entered (typing the space bar or pressing a key too lightly to activate the keyboard).

The first type of error was considered a "legitimate" error inasmuch as the examinee responded with one of the anticipated alternative answers. The blank and "adjacency" errors were considered to be more serious, suggesting an affect of technology on the scores, resulting in lower test reliability. The anlysis of these two types of errors concentrated on items 26 through 125 because corrective feedback was given to examinees on the earlier items.

TABLE 4

ERROR FREQUENCIES FOR STIMULUS CONDITIONS AND TYPING SKILL LEVEL

Type of Error	Condition	Number of Errors	Number of Examinees Making Errors
Adjacent	Letter	10	6
	Number	28	13
Blank	Letter	5	5
	Number	27	7
Adjacent	Typists	15	7
	Non Typists	23	12
Blank	Typists	8	5
	Non Typists	24	7

Table 4 reports the number of errors and the number of examinees committing an error for each of the two kinds of "illegitimate" errors. It can be seen in Table 4 that adjacent errors were made frequently by nontypists and under the number condition. This was the case both for total number of errors made and for total number of examinees making errors. Nontypists and the number condition accounted for a large number of blank responses as well.

DISCUSSION

The motivation for this investigation was a belief that computer-based testing has a potential value in education but that little evidence had been reported as to the nature of test reliability, as expressed by response accuracy, under computer administration. Related studies dealt with paper and pencil tests or were more related to digit span and short term memory. The present study investigated response accuracy on a computer-administered test.

The conclusions which may be reached as a result of this study must be qualified by the sample, which yielded a disproportionate number of nontypists in the N-group. One might expect nontypists to perform more poorly than typists on a typewriter keyboard. Although the nontypists group had higher variance, significant differences between the typists and nontypist groups were not shown. Even though the nontypist group did not have signicantly lower scores, the number-group (which contained a large number of nontypist participants) did perform less well. This finding may tentatively suggest that number response alternatives not be used on computer-based multiple choice tests. The nature of the errors, as indicated in Table 4, also suggests that number-responses introduce errors with higher frequency than letter-responses. It is important to note that the number of examinees who made errors was comparable for the two groups. Those examinees who did make errors did so with higher frequency under the number condition.

The study described is limited in generalizibility as the items showed no face validity as items. The study should be replicated with more realistic-looking items and with a more balanced sample. Efforts are underway to correct these deficiencies.

REFERENCE

Cashen, V.M., & Remseyer, G. The use of separate answer sheets by primary age children. *Journal of Educational Measurement*, 1969, *6*(3), 155-158

Conrad, R., & Hull, A.J. The preferred layout for numeral data - entry keysets. *Ergonomics*,1968 *11*(2), 165-173

Glass, G.V., & Stanley, J.C. *Statistical Methods in Education and Psychology*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

Miller, I., & Minor, F.J. Influence of multiple choice answer form design on anwer-marking performance. *Journal of Applied Psychology*, 1963, *47*(6), 374-379

Muller, D., Calhoun, E., & Orling, R. Test reliability as a function of answer sheet mode. *Journal of Educational Measurement*, 1972, *9*(4), 321-324

Winer, B.J., *Statistical Principles in Experimental Design* (2nd ed.). New York: McGraw-Hill, 1971.